titanic_data

사용 데이터 : 타이타닉(titanic)

타이타닉 배에 있던 승객들 명단

데이터 출처| seaborn에 있는 데이터 중 titanic 선택

데이터 colum 의미

survived : 생존여부 / 0=사망, 1=생존

pclass : 좌석등급 / 1=1등석, 2=2등석, 3=3등석

sex : 성별 / male=남, female=여

age : 나이

sibsp : 동승한 자매,배우자 수

parch : 동승한 부모,자식 수

fare : 요금

embarked : 탑승지 / C=셰르부르, Q=퀸즈타운, S=사우샘프턴

class : 좌석등급 / First, Second, Third

who : 성별 / man, woman

adult_male : 성인 남성 여부

deck : 선실 고유 번호 가장 앞자리 알파벳 / A,B,C,D,E,F,G

embark_town : 탑승지 명 / Cherbourg, Queenstown, Southampton

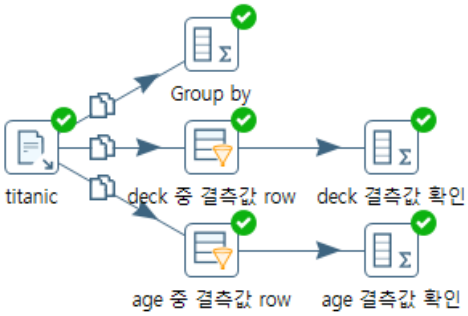
alive : 생존여부

alone : 혼자 여부

타이타닉 데이터

#	survived	pclass	sex	age	sibsp	parch	fare	embarked	class	who	adult_male	deck	embark_town	alive
1	0	3	male	22.0	1	0	7.25	S	Third	man	true	<null>	Southampton	no
2	1	1	female	38.0	1	0	71.2833	C	First	woman	false	C	Cherbourg	yes
3	1	3	female	26.0	0	0	7.925	S	Third	woman	false	<null>	Southampton	yes
4	1	1	female	35.0	1	0	53.1	S	First	woman	false	C	Southampton	yes
5	0	3	male	35.0	0	0	8.05	S	Third	man	true	<null>	Southampton	no
6	0	3	male	<null>	0	0	8.4583	Q	Third	man	true	<null>	Queenstown	no
7	0	1	male	54.0	0	0	51.8625	S	First	man	true	E	Southampton	no
8	0	3	male	2.0	3	1	21.075	S	Third	child	false	<null>	Southampton	no
9	1	3	female	27.0	0	2	11.1333	S	Third	woman	false	<null>	Southampton	yes
10	1	2	female	14.0	1	0	30.0708	C	Second	child	false	<null>	Cherbourg	yes
11	1	3	female	4.0	1	1	16.7	S	Third	child	false	G	Southampton	yes
12	1	1	female	58.0	0	0	26.55	S	First	woman	false	C	Southampton	yes
13	0	3	male	20.0	0	0	8.05	S	Third	man	true	<null>	Southampton	no
14	0	3	male	39.0	1	5	31.275	S	Third	man	true	<null>	Southampton	no
15	0	3	female	14.0	0	0	7.8542	S	Third	child	false	<null>	Southampton	no
16	1	2	female	55.0	0	0	16.0	S	Second	woman	false	<null>	Southampton	yes
17	0	3	male	2.0	4	1	29.125	Q	Third	child	false	<null>	Queenstown	no
18	1	2	male	<null>	0	0	13.0	S	Second	man	true	<null>	Southampton	yes
19	0	3	female	31.0	1	0	18.0	S	Third	woman	false	<null>	Southampton	no
20	1	3	female	<null>	0	0	7.225	C	Third	woman	false	<null>	Cherbourg	yes

결측값 확인하기



총 row 개수는 **891**개

#	총_row_개수
1	891

그 중 **deck**의 결측치 개수 : **688**개

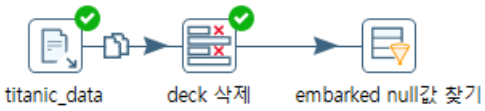
#	deck	deck_결측치_개수
1	<null>	688

age의 결측치 개수 : **177**개

#	age	age_결측치_개수
1	<null>	177

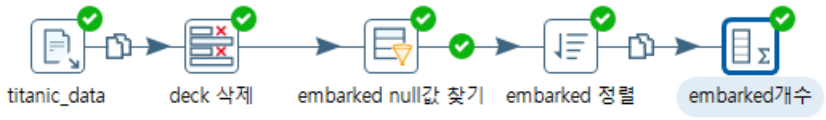
column deck 삭제

embarked null값인 row찾기



#	survived	pclass	sex	age	sibsp	parch	fare	embarked	class	who	adult_male	embark_town	alive	alone
1	1	1	female	38.0	0	0	80.0	<null>	First	woman	false	<null>	yes	true
2	1	1	female	62.0	0	0	80.0	<null>	First	woman	false	<null>	yes	true

1번 null값



#	embarked	embarked_개수
1	<null>	2
2	C	17
3	S	14

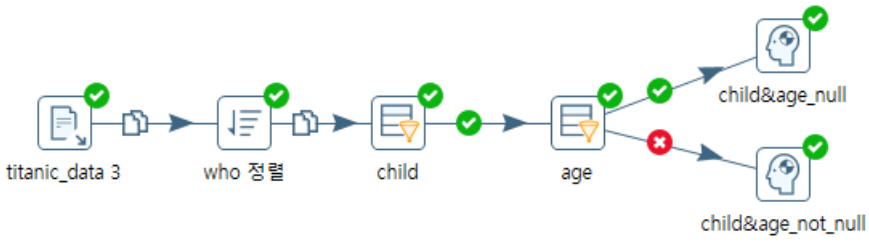
`성별:female, 좌석등급:first, 생존여부:생존, 혼자여부:혼자`인 사람들의 embarked별 개수 파악 → 가장 많은 C로 변경

C, Cherbourg 으로 변경

age null값인 row찾기

`성별, 생존여부, 좌석등급, 혼자여부`인 사람들의 age별 개수 파악 → **각 그룹별 age 평균**으로 null값 대체

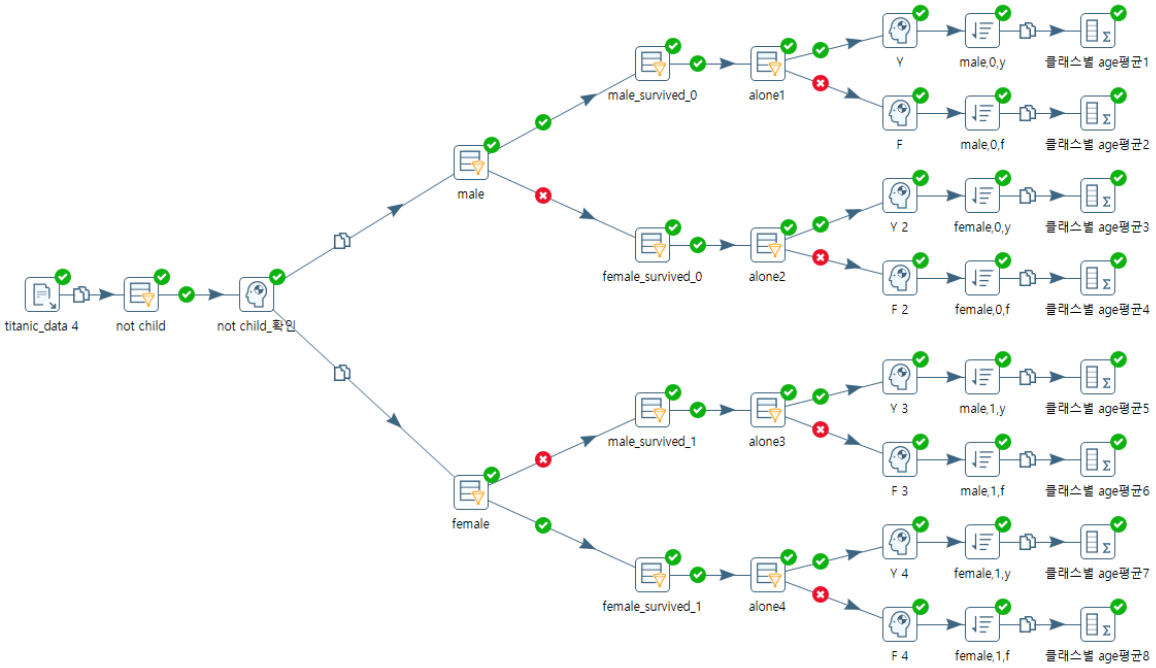
1. **age에 영향을 줄수있는 나이대인 who=child에는 null값이 없다는 것 확인**



Filter rows를 통해 who=child로 필터링 후, age에서 null값 있는지 확인 → child의 age에는 null값이 없음

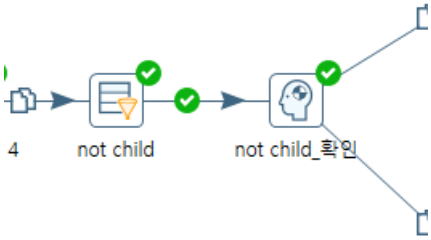
2. **그룹별 나이 평균 알아보기**

그룹 나누는 조건을 여러개 설정해서 그룹이 총 8개(+각 그룹당 3개의 좌석 등급 → 총 24개)로 많다

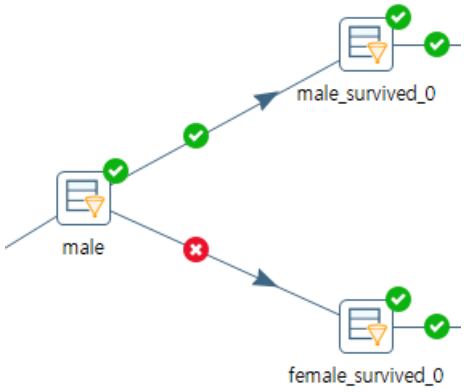


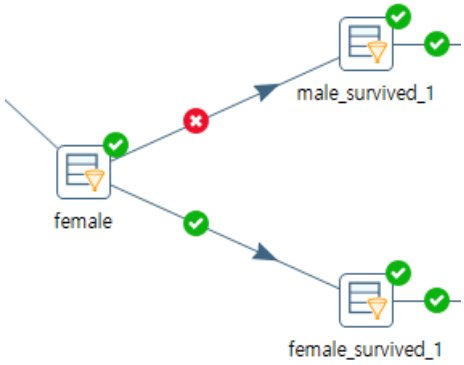
+ 상세

▼ who=child에는 결측값이 없으므로 man, woman에서 확인하기

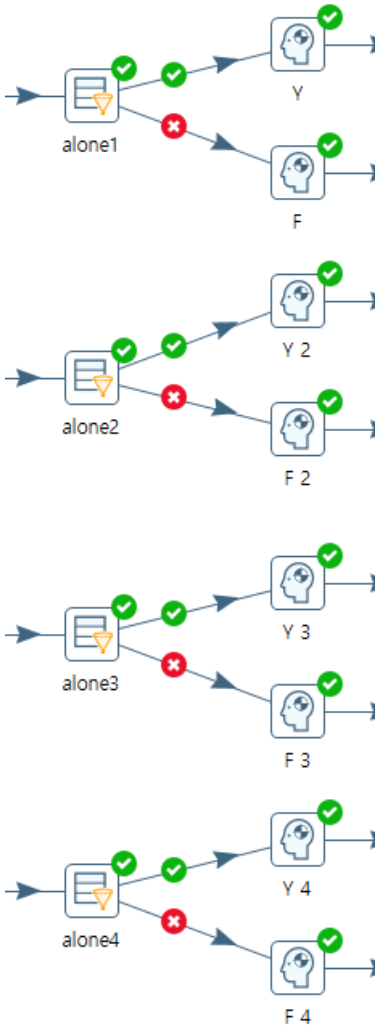


▼ 성별 male, female과 생존 여부로 1차적으로 나누기

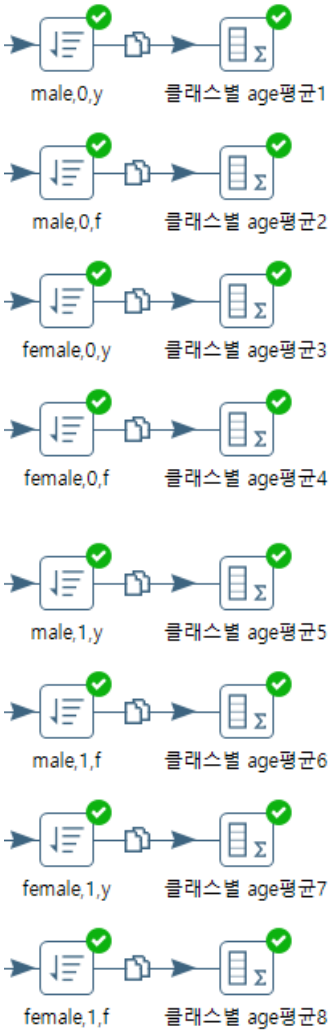




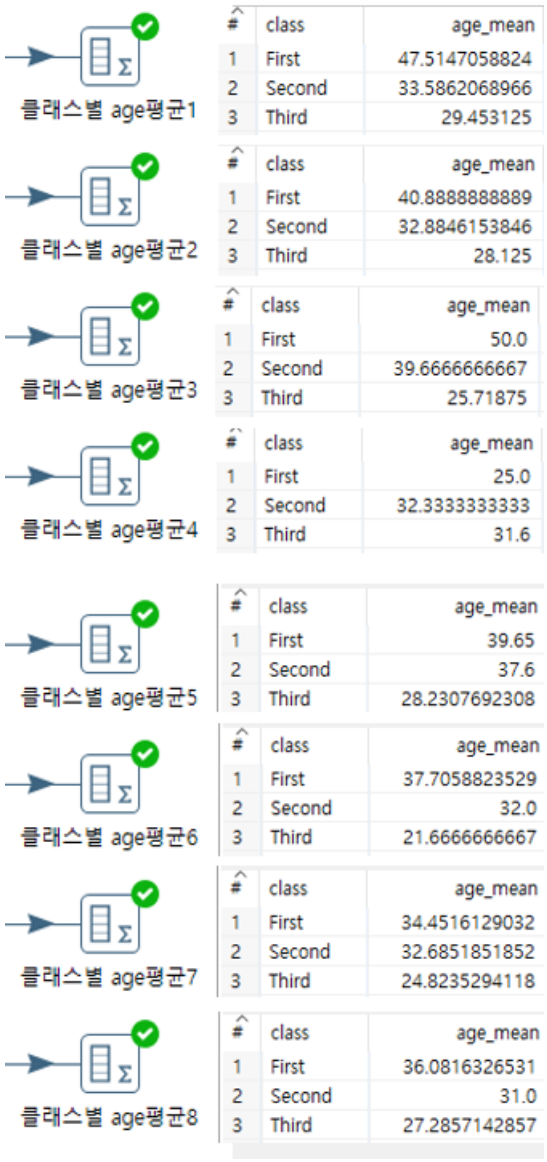
▼ 혼자왔는지 여부로 2차 나누기



▼ 최종적으로 클래스별 나이 평균 구하기


male,0,y
클래스별 age평균1
male,0,f
클래스별 age평균2
female,0,y
클래스별 age평균3
female,0,f
클래스별 age평균4
male,1,y
클래스별 age평균5
male,1,f
클래스별 age평균6
female,1,y
클래스별 age평균7
female,1,f
클래스별 age평균8

결과



#	class	age_mean
1	First	47.5147058824
2	Second	33.5862068966
3	Third	29.453125

#	class	age_mean
1	First	40.8888888889
2	Second	32.8846153846
3	Third	28.125

#	class	age_mean
1	First	50.0
2	Second	39.6666666667
3	Third	25.71875

#	class	age_mean
1	First	25.0
2	Second	32.3333333333
3	Third	31.6

#	class	age_mean
1	First	39.65
2	Second	37.6
3	Third	28.2307692308

#	class	age_mean
1	First	37.7058823529
2	Second	32.0
3	Third	21.6666666667

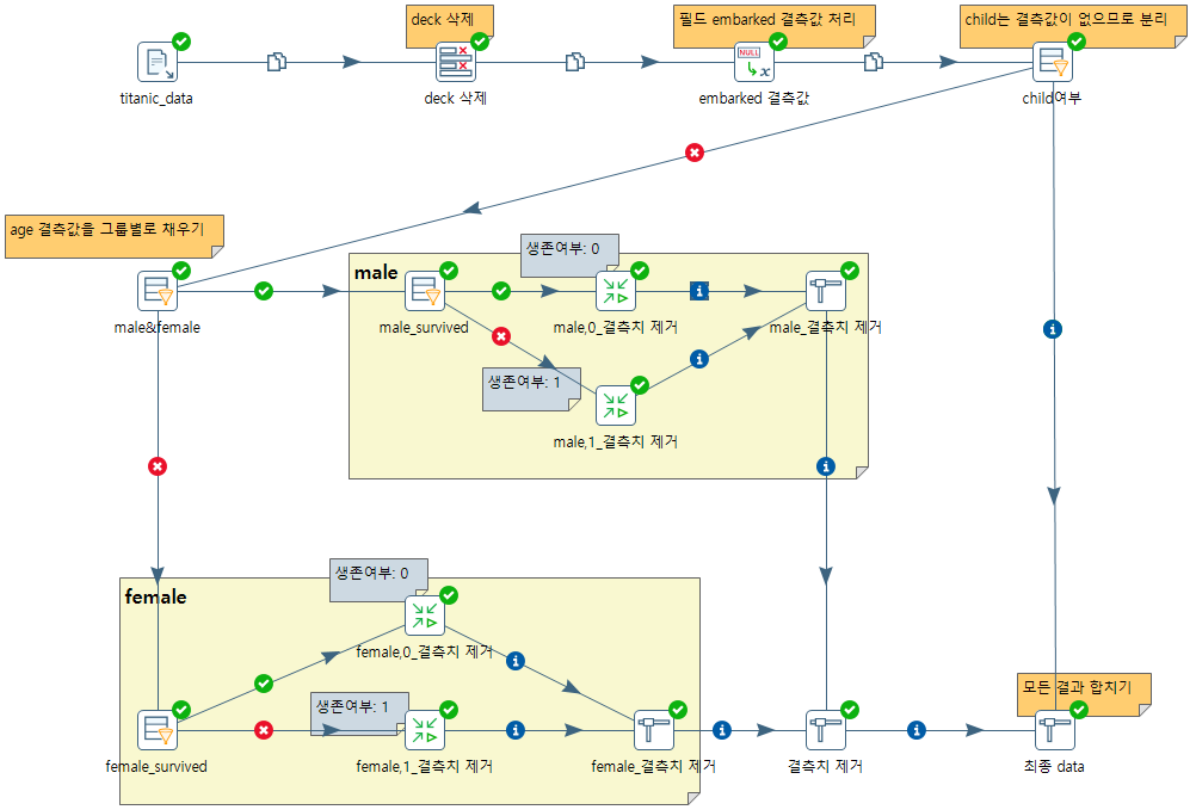
#	class	age_mean
1	First	34.4516129032
2	Second	32.6851851852
3	Third	24.8235294118

#	class	age_mean
1	First	36.0816326531
2	Second	31.0
3	Third	27.2857142857

전체적으로 등급이 아래일 수록 탑승객의 나이가 어리다는 것도 알 수 있다

최종, 컬럼 삭제 및 null 채우기

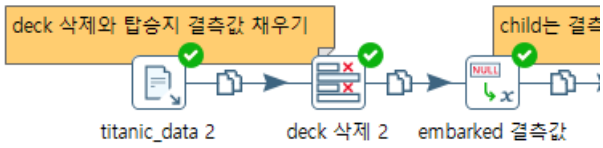
각 그룹별 평균 나이는 반올림으로 채우기



+ 상세

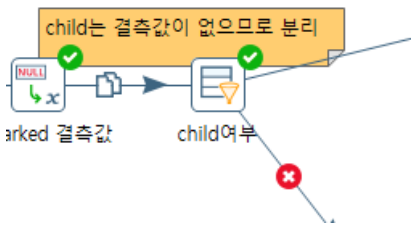
▼ column deck 삭제 & embarked 결측값 대체하기

column deck은 삭제하고, embarked는 C로 대체한다.



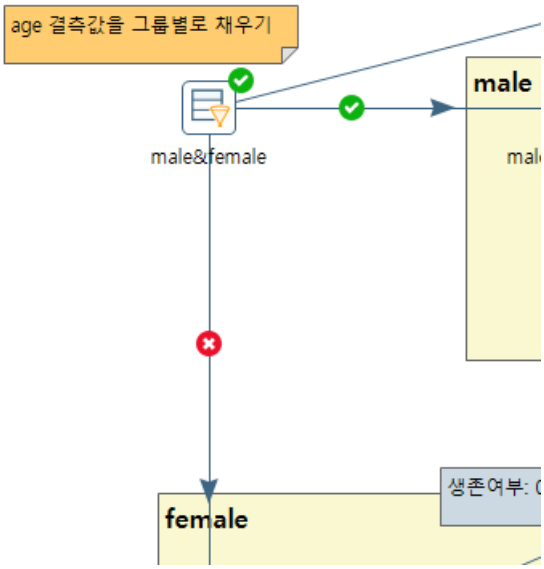
▼ 성인과 아이 구분

child에는 결측값이 없으므로, Filter rows를 통해 who가 child면 True, 아닐경우 False가 되도록 설정한다.

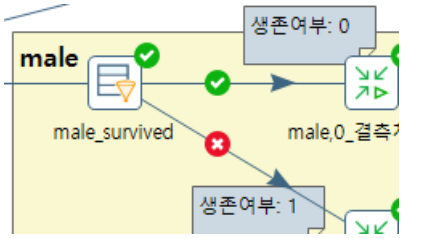


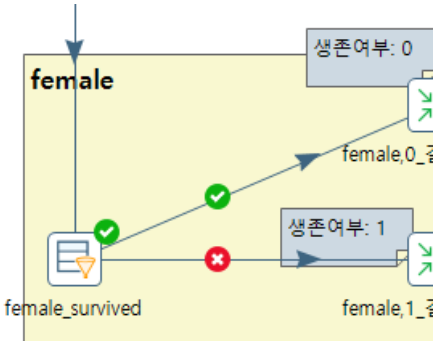
▼ 남성과 여성 구분 & 생존 여부 구분

성인 중 남성과 여성을 구분한다.



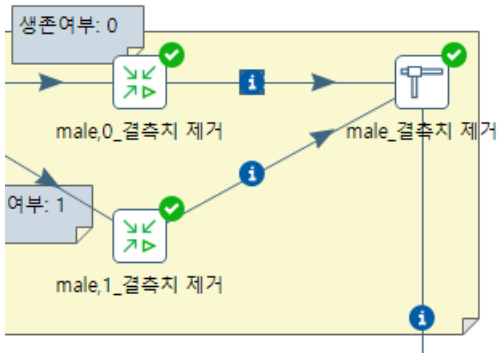
남성,여성 별로 생존여부로 구분한다.

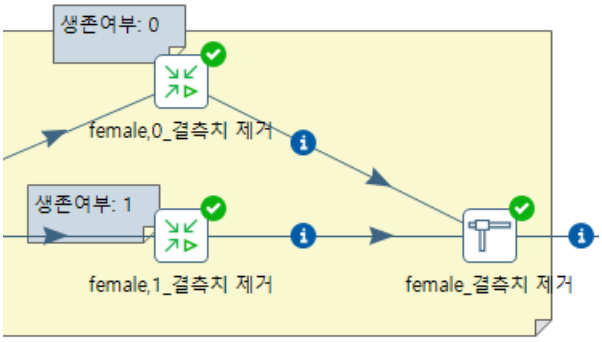




▼ 혼자여부와 좌석 등급 별 나이 평균 값으로 결측값 대체하기

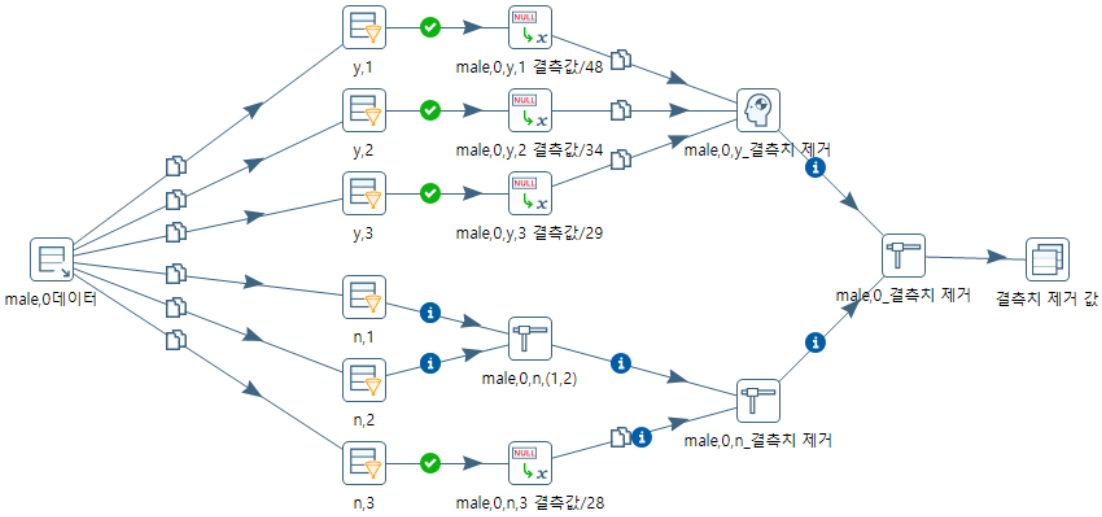
혼자 여부와 좌석 등급별로 나눠 각 null값을 각 그룹의 age 평균으로 대체해준다



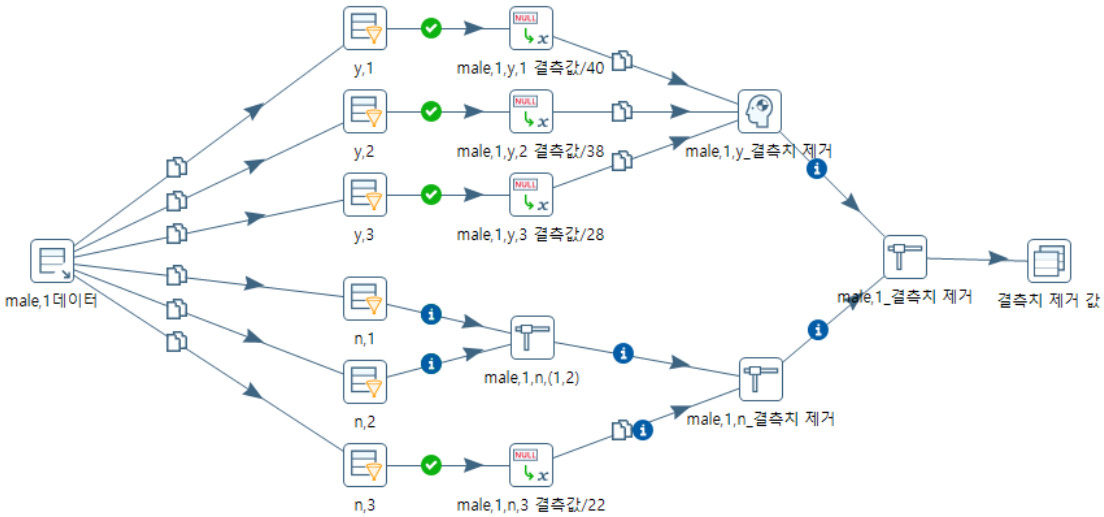


그룹 중 결측치가 없는 그룹은 별다른 처리 없이 합쳐주었다.

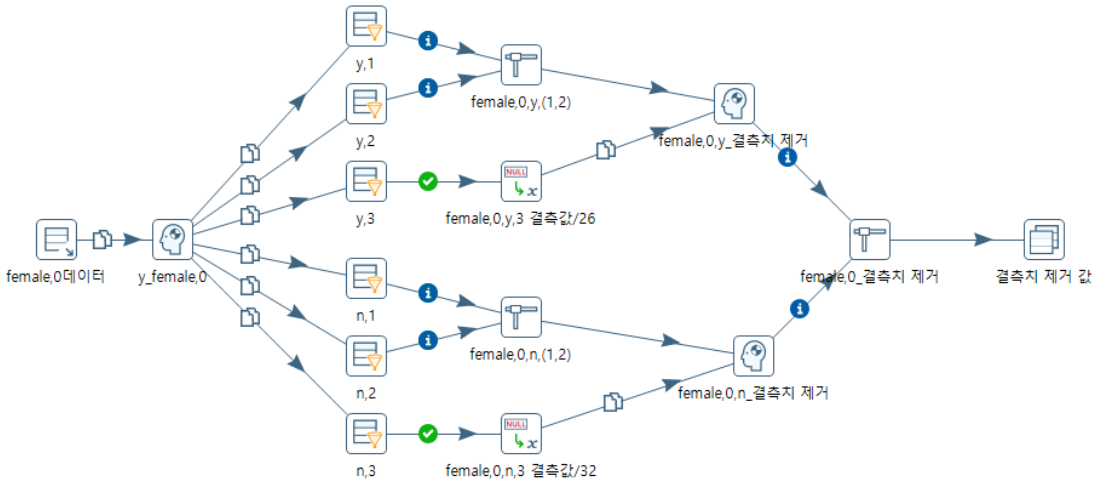
▼ male,0_결측치 제거 step



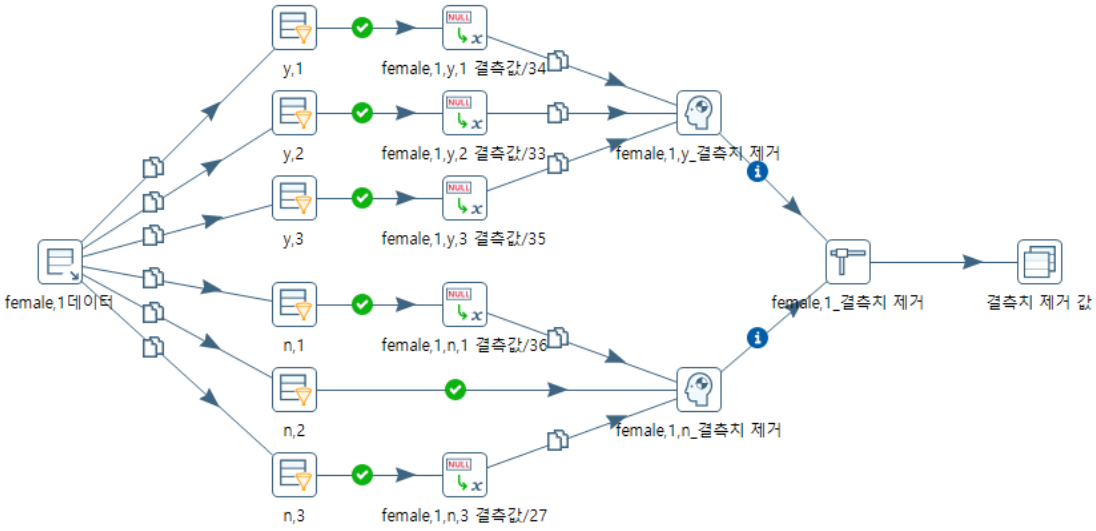
▼ male,1_결측치 제거 step



▼ female,0_결측치 제거 step

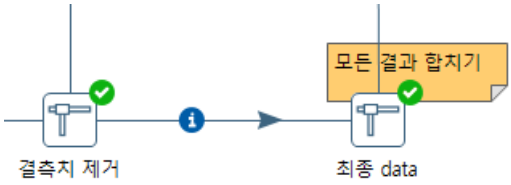


▼ female,1_결측치 제거 step

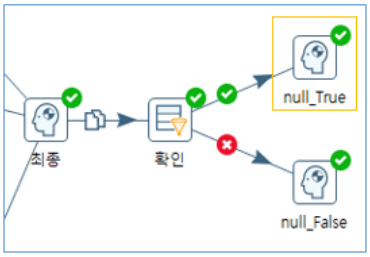


▼ 결측치가 처리된 값들 모두 합치기

최종적으로 모든 값들을 합쳐준다.



null값 처리가 잘 됐는지 확인 작업

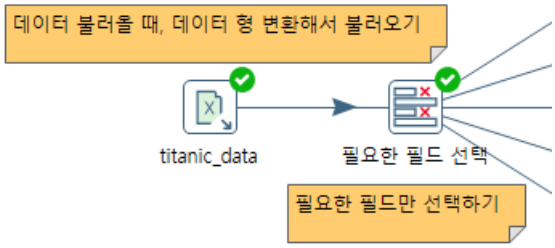


#	survived	pclass	sex	age	sibsp	parch	fare	embarked	class	who	adult_male	embark_town	alive	alone
1														

결측값 처리한 data 살펴보기

생존여부 비율 살펴보기

앞 부분은 모두 동일 (전처리한 데이터 불러오기 → 필드 선택)



생존비율을 통해 알 수 있는 내용

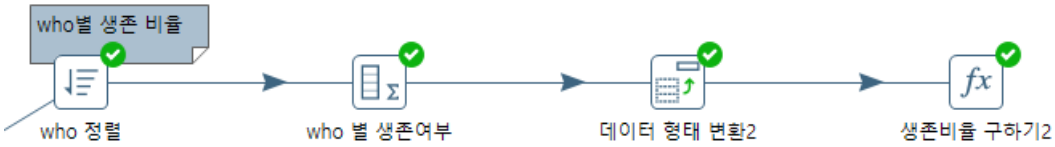
▼ 1. 남성, 여성 별 생존자 수 & 비율

#	sex	0	1	합	생존비율
1	female	81	233	314	74.2
2	male	468	109	577	18.89

결과

남성에 비해 여성의 생존 비율이 약 3.9배 높다.

▼ 2. 성인_남, 성인_여, 어린이 별 생존자 수 & 비율



#	who	0	1	합	생존비율
1	child	34	49	83	59.04
2	man	449	88	537	16.39
3	woman	66	205	271	75.65

결과

성인 여성의 생존 비율이 75.65%로 성인 남성에 비해 약 4.6배 높다.

어린이의 생존 비율은 59.04%이고,

탑승객 중 성인 남성의 수가 가장 많으며, 성인 여성의 2배, 어린이의 6.4배 많다.

▼ 3. 성인_남,여 / 어린이_남,여 별 생존자 수 & 비율

#	구분	0	1	합	생존비율
1	female_child	15	28	43	65.12
2	female_woman	66	205	271	75.65
3	male_child	19	21	40	52.5
4	male_man	449	88	537	16.39

결과

성인 남성, 성인 여성, 남자 어린이, 여자 어린이를 비교해보면,

성인 여성과 여자 어린이의 순으로 생존확률이 높으며

성인 남성의 생존률은 다른 그룹에 비해 현저히 적은 것을 볼 수 있다.

▼ 4. 좌석 등급 별 생존자 수 & 비율

#	pclass	0	1	합	생존비율
1	1	80	136	216	62.96
2	2	97	87	184	47.28
3	3	372	119	491	24.24

결과

1등급의 좌석에 탑승한 승객의 생존 비율이 62.96%로 가장 높으며, 좋은 등급의 좌석에 탔을 경우, 생존할 확률이 높았다.

▼ 5. 동승자 수 별 생존자 수 & 비율

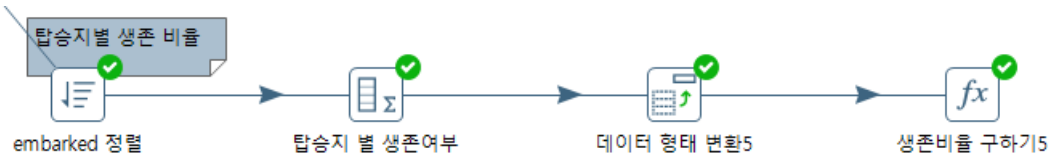
#	동승자	0	1	합	생존비율
1	0	374	163	537	30.35
2	1	72	89	161	55.28
3	2	43	59	102	57.84
4	3	8	21	29	72.41
5	4	12	3	15	20
6	5	19	3	22	13.64
7	6	8	4	12	33.33
8	7	6	0	6	0
9	10	7	0	7	0

결과

동승자 수가 3명일 경우의 생존율이 높았고, 그 다음으로는 2명,1명일 경우 순으로 생존률이 높았다.

동승자가 아에 없거나 너무 많은 경우, 생존률이 적다.

▼ 6. 탑승지 별 생존자 수 & 비율

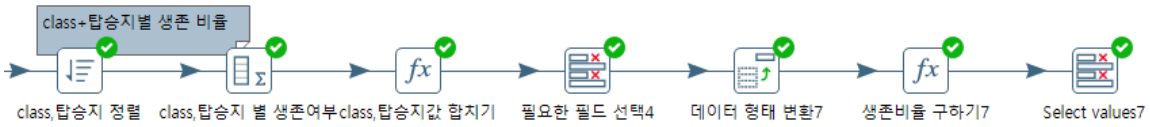


#	embarked	0	1	합	생존비율
1	C	75	95	170	55.88
2	Q	47	30	77	38.96
3	S	427	217	644	33.7

결과

C, Cherbourg에서 탑승한 경우의 생존 비율이 가장 높았다.

▼ 7. 좌석등급&탑승지 별 생존자 수 & 비율



#	구분	0	1	합	생존비율
1	1_C	26	61	87	70.11
2	1_Q	1	1	2	50
3	1_S	53	74	127	58.27
4	2_C	8	9	17	52.94
5	2_Q	1	2	3	66.67
6	2_S	88	76	164	46.34
7	3_C	41	25	66	37.88
8	3_Q	45	27	72	37.5
9	3_S	286	67	353	18.98

결과

탑승지와 좌석 등급을 같이 보면 3등급 좌석을 이용한 고객의 탑승지가 S, Southampton인 경우가 많은 것을 볼 수 있다.

생존확률이 적은 3등급 좌석의 승객들의 탑승지가 S, Southampton인 경우가 많아 S에서 탑승한 승객의 생존률이 가장 적은 것으로 보인다.
